Genes & Diseases (2020) 7, 567–577

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://ees.elsevier.com/gendis/default.asp

FULL LENGTH ARTICLE

Amino acid variation analysis of surface spike glycoprotein at 614 in SARS-CoV-2 strains

Canhui Cao [a], Liang Huang [b], Kui Liu [c], Ke Ma [d], Yuan Tian [a,e], Yu Qin [a,e], Haiyin Sun [a,e], Wencheng Ding [a,e], Lingli Gui [f,*], Peng Wu [a,e,**]

[a] *Cancer Biology Research Center (Key Laboratory of the Ministry of Education), Tongji Medical College, Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China*
[b] *Department of Hematology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*
[c] *Department of Respiratory and Critical Care Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*
[d] *Department of Infectious Diseases, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*
[e] *Department of Gynecologic Oncology, Tongji Hospital, Tongji Medical College, Huazhong, China*
[f] *Department of Anesthesiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China*

Received 31 March 2020; received in revised form 1 May 2020; accepted 24 May 2020
Available online 2 June 2020

KEYWORDS
ACE2;
COVID-19;
Phylogenetic tree;
SARS-CoV-2;
Surface spike glycoprotein

Abstract As severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) continues to disperse globally with worrisome speed, identifying amino acid variations in the virus could help to understand the characteristics of it. Here, we studied 489 SARS-CoV-2 genomes obtained from 32 countries from the Nextstrain database and performed phylogenetic tree analysis by clade, country, and genotype of the surface spike glycoprotein (S protein) at site 614. We found that virus strains from mainland China were mostly distributed in Clade B and Clade undefined in the phylogenetic tree, with very few found in Clade A. In contrast, Clades A2 (one case) and A2a (112 cases) predominantly contained strains from European regions. Moreover, Clades A2 and A2a differed significantly from those of mainland China in age of infected population ($P = 0.0071$, mean age 40.24 to 46.66), although such differences did not exist between the US and mainland China. Further analysis demonstrated that the variation of the S

* Corresponding author.
** Corresponding author. Cancer Biology Research Center (Key Laboratory of the Ministry of Education), Tongji Medical College, Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China.
E-mail addresses: gui_lingli@hotmail.com (L. Gui), pengwu8626@tjh.tjmu.edu.cn (P. Wu).
Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2020.05.006

protein at site 614 (QHD43416.1: p.614D>G) was a characteristic of stains in Clades A2 and A2a. Importantly, this variation was predicted to have neutral or benign effects on the function of the S protein. In addition, global quality estimates and 3D protein structures tended to be different between the two S proteins. In summary, we identified different genomic epidemiology among SARS-CoV-2 strains in different clades, especially in an amino acid variation of the S protein at 614, revealing potential viral genome divergence in SARS-CoV-2 strains.

Introduction

In late December 2019, an unknown pneumonia-like disease linked to a novel coronavirus was first identified. The disease, now named Coronavirus Disease 2019 (COVID-19), was found to be caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1] By 27—28 February 2020, the virus had caused more than 82 000 infections and 2 800 deaths[2] and had spread to more than 50 countries.[3] By 11 March 2020, global infections reached 118 000, with 4 291 deaths across 114 countries, prompting the World Health Organization (WHO) Chief, Tedros Adhanom Ghebreyesus, to declare COVID-19 a global pandemic. As of 30 April 2020, infections reached 3 090 445, including 217 769 deaths.[4] These figures highlight the worrisome speed with which this outbreak has spread.

Many SARS-related coronaviruses (SARSr-CoVs) have been identified in bats, which are considered natural reservoir hosts.[5—8] Accumulating evidence has shown that the genomic sequence of SARS-CoV-2 is similar to that of SARS-CoV and thus belongs to the SARSr-CoV family. At the whole-genome level, the new virus shares about 96% identity to a known bat coronavirus (BatCoV RaTG13).[9] Furthermore, SARS-CoV-2 uses the same cell entry receptor as SARS-CoV via angiotensin-converting enzyme II (ACE2),[1,9] which is associated with the spread of SARS-CoV-2 among the human population.[10] In SARS-CoV infection, the surface spike glycoprotein (S protein) mediates receptor recognition with ACE2 and membrane fusion conformation to facilitate entry of the virus into the host cell.[11,12]

The rapid spread of COVID-19, which has led to a global pandemic, has placed public health systems under severe pressure. According to prospective research on infected patients, scientists have found that patients infected with COVID-19 had high amounts of proinflammatory cytokines (IL1B, IFNγ, IP10, and MCP1), probably leading to the responses of activated T-helper-1 (Th1) cells.[13] However, the pathogenic mechanism of the disease and reason for its rapid spread remain unclear. Haplotype analyses based on genome variations have also demonstrated viral genome divergence among strains from different regions and countries.[14] In addition, a group of 27 public health scientists have strongly condemned the rumors that SARS-CoV-2 is not of natural origin.[15] Like other coronaviruses, SARS-CoV-2 appears to accumulate, on average, one or two mutations per month.[16] Thus, identifying variations in strains from different regions is a key factor for understanding the pathogenic mechanisms of this disease.

In this study, we explored the genomes of 489 SARS-CoV-2 strains derived from 88 regions in 32 countries between December 2019 and March 2020 via the Nextstrain database of viral genomes.[17,18] We found that Clades A2 and A2a in the downloaded phylogenetic tree contained strains mostly from Europe and several South American countries. Further analysis demonstrated that amino acid variation of the S protein at 614 (QHD43416.1: p.614D>G), i.e., substitution of glutamic acid (D) with glycine (G) in the mutant protein, was found in strains within Clades A2 and A2a. Furthermore, we used computational methods based on evolutionary principles to predict the effect of such variation on protein function. These results provide important insights into amino acid variation of the S protein at site 614 in SARS-CoV-2, hinting at potential viral genome divergence in SARS-CoV-2 strains.

Materials and methods

Phylogenetic tree of SARS-CoV-2 in Nextstrain

The phylogenetic tree of SARS-CoV-2 was performed in Nextstrain (https://nextstrain.org/ncov),[17,18] using the parameter with clade, countries, and genotype of surface spike glycoprotein (S protein) at site 614 with default settings, to identify the genomic epidemiology of COIVD-19. SARS-CoV-2 was defined as "Novel coronavirus (2019-nCoV)" in the dataset, which included 489 SARS-CoV-2 genomes from 88 regions of 32 countries between December 2019 and March 2020.

The world map of the SARS-CoV-2

The world map of the SARS-CoV-2 relating clade was performed in Nextstrain (https://nextstrain.org/ncov).[17,18] Geographic resolution used division as a parameter. The map, pie chart, and transmissions were saved with default settings.

Surface spike glycoprotein (S protein) amino acid sequence

Surface spike glycoprotein (S protein) amino acid sequence of SARS-CoV-2 was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/protein/QHD43416.1), with accession MN908947.3, 1273 aa.[19] The amino acid variation of S protein at site 614, with QHD43416.1: p.614D>G. In the gene

of spike glycoprotein at c.1841 with gene-S: c.1841gAt > gGt.

Predicting coding nonsynonymous variation effect on protein function

The variation of S protein at 614 (QHD43416.1: p.614D>G) was predicted to affect S protein function via PROVEAN (v1.1) (http://provean.jcvi.org/index.php)[20,21] and PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/index.shtml).[22] The sequence of QHD43416.1 and the amino acid variation site 614 were used in Protein Variation Effect Analyzer (PROVEAN (v1.1)) with the default threshold: 2.5. The words "does not affect," and "does not reduce" were classified as a neutral effect. PolyPhen-2 also used the sequence of QHD43416.1 and p.614D>G site information. The closer the score is to 1.0, the greater the effect on protein function, and the closer the score is to 0, the smaller the effect on protein function. The human divergence (HumDiv) dataset is a preferred model for evaluating rare alleles, mapping dense regions, and analyzing natural selection. The human variation (HumVar) dataset is a preferred model for diagnosis of Mendelian disease, including many mildly deleterious alleles.

Homology modeling of S protein

Homology modeling of the S protein (QHD43416.1) and the variation of S protein (QHD43416.1: p.614D>G) were built in SWISS-MODEL Server (https://swissmodel.expasy.org/interactive).[23,24] The only different sequence between the two proteins was the site 614, from glutamic acid (D) to glycine (G). The input format was FASTA that was downloaded from NCBI.

Protein modeling estimate

Homology modeling protein of S protein (QHD43416.1) and the variation of S protein (QHD43416.1: p.614D>G) of SARS-CoV-2 were estimated via SWISS-MODEL Server (22, 23) (https://swissmodel.expasy.org/interactive). Sequence identities with surface glycoprotein were 99.26% and 99.17%, respectively. The comparison included local quality estimate (LQE) and global quality estimate (GQE) via QMEAN. The LQE represented ensemble information from all template structures found.

Protein structural visualization

Protein structural visualization was performed in SWISS-MODEL Server via NGL (WebGL), implementing using PV (https://biasmv.github.io/pv/), an interactive WebGL viewer based on 3D protein structure. Protein three-dimensional structure displayed with Spacefill. Model-template alignment used 3 model protein and query (QHD43416.1 or QHD43416.1: p.614D>G). p.614D>G was zoomed in the visualization of the web server.

Statistical analysis

Data were presented as the number of infections in age 1–20, 21–60, >60; mean age of country or Clade A2a; the number of females or males. Results were calculated by GraphPad Prism 6 (version 6.02) software and interpreted by the Mann–Whitney U test or Kruskal–Wallis test as indicated in the legends. Bold *P* values indicate $P < 0.05$.

Results

Different SARS-CoV-2 clades among countries in phylogenetic tree

A phylogenetic tree of nucleotide sequences can be used to compare genome divergence and similarity among viruses.[19,25,26] Here, we performed phylogenetic tree analysis based on the consolidation of 489 genomes of SARS-CoV-2 strains sampled from 88 regions of 32 countries between December 2019 and March 2020 from the Nextstrain database (Fig. 1A). Results showed that the genomes of strains from mainland China were mostly distributed in Clade B and Clade undefined in the phylogenetic tree, with only 3.47% (5/144) found in Clade A. In addition, Clades A2 (one case) and A2a (112 cases) contained no cases from mainland China. In these clades, all cases came from 16 regions, mainly the Netherlands (65 cases), Switzerland (13 cases), and UK (13 cases), with only one case reported from Taiwan (Table 1).

Furthermore, Clades A2 and A2a differed from mainland China in regard to age of infected population ($P = 0.0071$, mean age 40.24 to 46.66), which was also observed between the UK and China ($P = 0.0102$, mean age 46.17 to 46.66). However, this difference was not found between the US and China or between Switzerland and China (Table 2). In addition, based on the map of infection clades worldwide (Fig. 1B), genomes from mainland China strains were mostly found in Clades B, B1, and B2, which differed from the clades containing strains from Europe and several South American countries.

Amino acid variation of S protein at site 614 in Clade A2 and A2a SARS-Cov-2 strains

It is now known that SARS-CoV-2 uses the same cell entry receptor as SARS-CoV, i.e., in infection, the S protein mediates receptor recognition with ACE2 and membrane fusion conformation to facilitate virus entry into the host cell.[1,9] Thus, the S protein in SARS-CoV is a target for the development of vaccines.[27] Here, we performed phylogenetic tree analysis by genotyping the S protein in SARS-CoV-2 strains. We found amino acid variation of the S protein at site 614 (i.e., substitution of glutamic acid (D) with glycine (G) (QHD43416.1: p.614D>G)) in strains from Clades A2 and A2a (Fig. 2A and B). Further analysis showed that diversity between the S protein and mutant S protein at site 614 (i.e., genome sequence from 23 402 to 23 404) was 0.598% (Fig. 2C). We also found other amino acid variations in the S protein of SARS-CoV-2 strains, including at site 49 (p.49H > Y) and 1044 (p.1044G). However, the frequency of

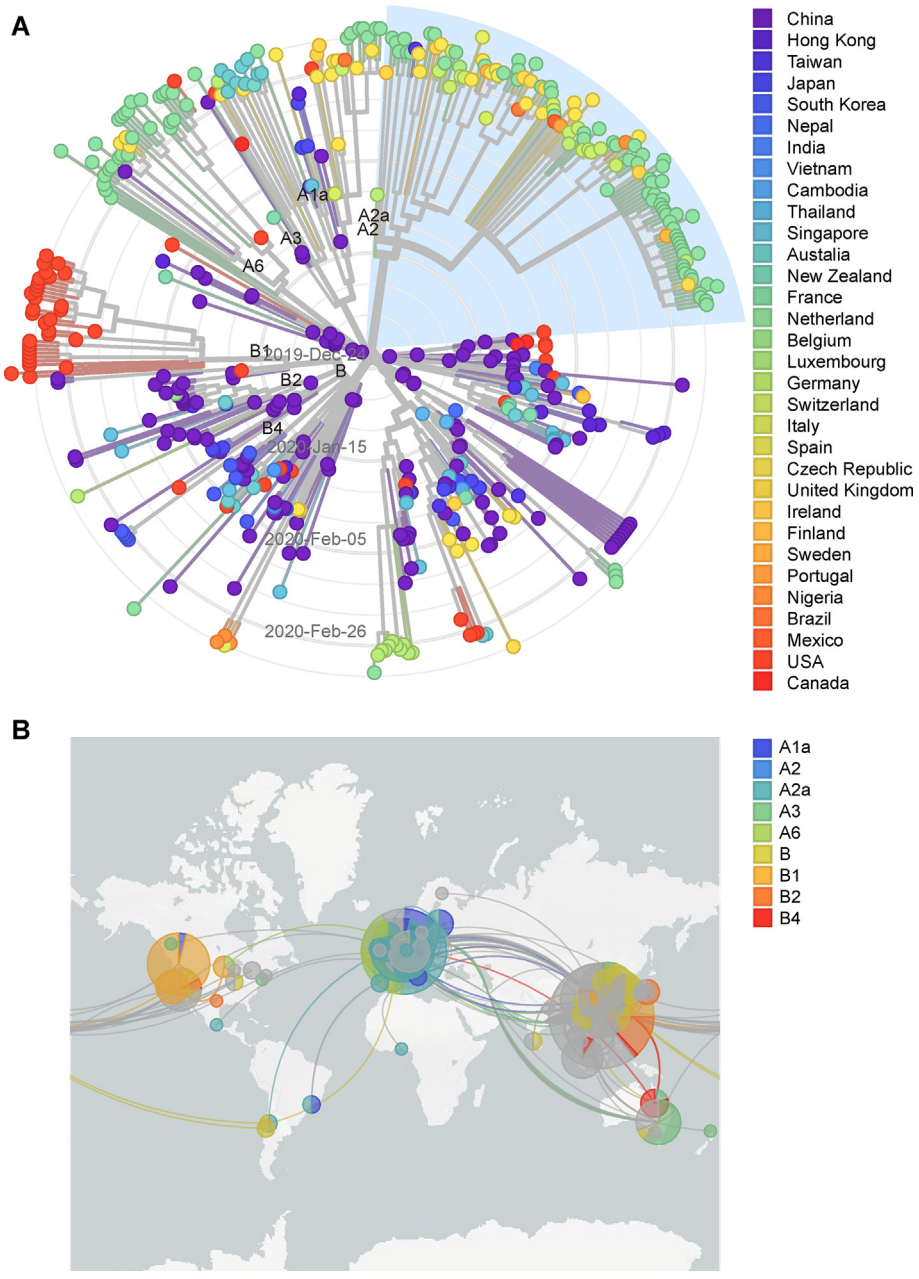


Figure 1 Different SARS-CoV-2 clades among countries in phylogenetic tree. (A) Phylogenetic tree of 489 SARS-CoV-2 genomes from Nextstrain, the cases were colored by countries. Branch labels were clades. **(B)** Clade Distribution of 489 SARS-CoV-2 genomes in world map from Nextstrain. Color by clades.

these variations was low (Fig. 2C), and the strains were sporadically reported in different regions. Based on whole-genome sequence data of SARS-CoV-2 strains, the amino acid variation of the S protein at site 614 was not found in the strains from the B and undefined clades.

Variation in S protein at site 614 does not affect protein function

Nonsynonymous mutations contribute to phenotypic differences in the human population, as well as susceptibility to genetic diseases[28] and functional diversity of proteins.[29] Here, we applied computational methods based on evolutionary principles to predict the effect of S protein variation on its function. Using PROVEAN v1.1 (Protein Variation Effect Analyzer),[20,21] we obtained a PROVEAN score of 0.215 (Supplementary Table 1), with 758 supporting sequences for prediction (Supplementary Table 2). Results predicted that variation of the S protein at 614 would have a neutral effect on protein function. We also employed PolyPhen-2 [22] to annotate coding nonsynonymous variation. Results predicted that variation of the S protein at 614 would have a benign effect on protein function, with scores of 0.004 (sensitivity: 0.97; specificity: 0.59) and 0.012 (sensitivity: 0.96; specificity: 0.52) (Fig. 3A and B and

Table 1 Epidemiology characteristics of countries in Clades A2 and A2a.

Countries	Cases in A2/A2a	Mean age	Female number	Male number	Total cases	Percent in A2/Aa2
Netherlands	65	na	na	na	107	60.74%
Switzerland	13	33.31	2	11	14	92.86%
United Kingdom	13	46.27	4	7	33	39.39%
Ireland	4	28.5	2	2	5	80.00%
Finland	3	41	1	2	6	50.00%
Germany	3	na	na	1	15	20.00%
Italy	2	38	1	1	4	50.00%
Portugal	2	46.5	0	2	2	100.0%
Brazil	1	61	0	1	2	50.00%
Spain	1	56	0	1	2	50.00%
Luxembourg	1	na	na	na	1	100.0%
Mexico	1	35	0	1	1	100.0%
Nigeria	1	na	0	1	1	100.0%
Czech Republic	1	44	0	1	1	100.0%
Chile	1	40	1	na	4	25.00%
Taiwan	1	66	1	0	7	14.29%

Supplementary Table 3). These results demonstrate that variation of the S protein at 614 was not predicted to affect protein function.

Protein modeling estimate of QHD43416.1 and QHD43416.1: p.614D > G

We modeled the S protein (QHD43416.1) and mutant S protein (QHD43416.1: p.614D>G) using SWISS-MODEL,[23,24] speculating on the quality estimates of the two proteins. Amino acid variation was found at site 614, i.e., substitution of glutamic acid (D) with glycine (G) (Fig. 4A). After modeling, the sequence identities of the surface glycoproteins were 99.26% and 99.17%, respectively.

SWISS-MODEL relies on the QMEAN scoring system.[30] Here, the QMEAN values of the two proteins were −2,65 and −2.68; Cβ scores were −1.05 and −1.14; all-atom scores were −1.72 and −1.69; solvation scores were −1.11 and −1.22; and torsion scores were −2.02 and −1.99, respectively. However, local quality of the two proteins was similar (Fig. 4B).

Table 2 Epidemiology characteristics of China, Clade A2a, UK, Switzerland, USA.

	Countries/Clade				
Characteristics	China	A2a	UK	Switzerland	USA
Age					
1-20	8	2	0	0	0
21-60	63	36	27	13	15
>60	27	2	2	1	2
Mean age	46.66	40.24	46.17	33.29	49.06
P value		**0.0071**	**0.0102**	0.0974	0.1324
Sex					
Female	40	12	13	3	11
Male	69	30	16	11	8
P value		0.3463	0.4239	0.2594	0.0816

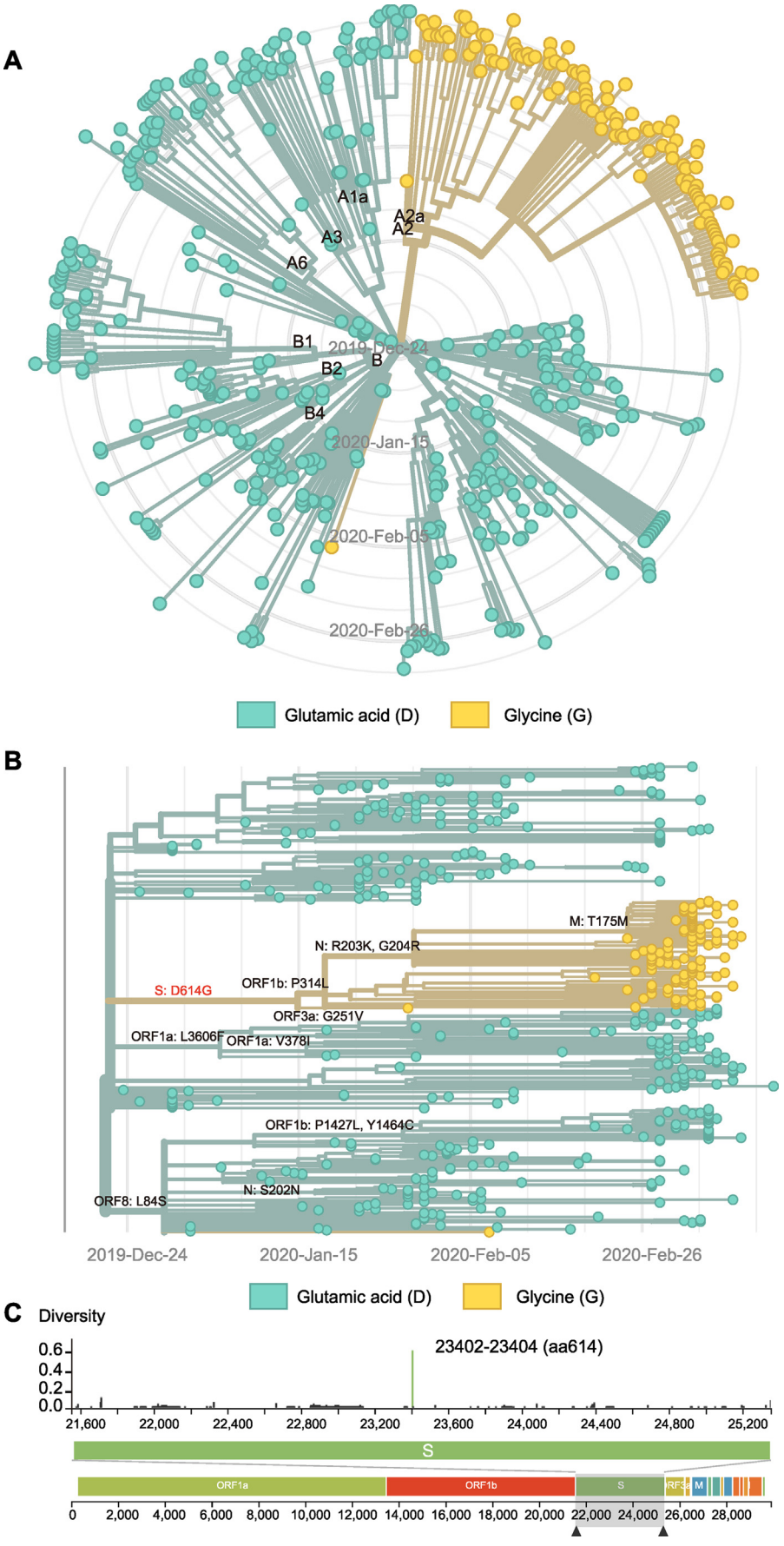
The epidemiology characteristics of A2a clade, UK, Switzerland, and USA were compared with those of China. *P*-value was calculated by the Mann–Whitney U test or Kruskal–Wallis test. Bold *P* values indicate $P < 0.05$. na: not applicable, meant the missing data.

Three-dimensional (3D) protein structure models of QHD43416.1 and QHD43416.1: p.614D > G

The 3D structure of a protein determines protein function and is uniquely established by the specificity of amino acid sequences.[31] Here, we used SWISS-MODEL[23,24] to predict the 3D protein structures of the S protein (QHD43416.1) and mutant S protein (QHD43416.1: p.614D>G) (Fig. 5A). In addition, we focused on site 614 of QHD43416.1 and QHD43416.1: p.614D>G (Fig. 5B). Importantly, glutamic acid (D) and glycine (G), marked in red in the two models, were in different positions of the 3D protein structure models.

Discussion

Due to continued growing number of infections regarding COVID-19, a pneumonia-like disease associated with SARS-CoV-2, the WHO declared the disease a public health emergency of international concern on 30 January 2020 and a global pandemic on 11 March 2020. Based on the genomic characterization of SARS-CoV-2 strains from 30 SARS-CoV-2 sequences, the disease outbreak was determined to have begun in mid-November 2019, and the genomes of strains provided clues about the viral diversity.[32] Thus, identifying variations in strains from different regions is a key factor for understanding the pathogenic mechanisms of this disease. By examining 489 SARS-CoV-2 genomes obtained from 32 countries sampled between December 2019 and March 2020 from the Nextstrain phylogenetic tree, we found that strains from mainland China were mostly distributed in Clade B and Clade undefined, and differed from strains in Clades A2 and A2a. Phylogenetic analysis allows researchers to compare the genome divergence between viral strains.


A
A1a
A2a
A2
A3
A6
B1
B2
B
B4
2019-Dec-24
2020-Jan-15
2020-Feb-05
2020-Feb-26
Glutamic acid (D)
Glycine (G)
B
M: T175M
N: R203K, G204R
S: D614G
ORF1b: P314L
ORF3a: G251V
ORF1a: L3606F
ORF1a: V378I
ORF1b: P1427L, Y1464C
N: S202N
ORF8: L84S
2019-Dec-24
2020-Jan-15
2020-Feb-05
2020-Feb-26
Glutamic acid (D)
Glycine (G)
C
Diversity
23402-23404 (aa614)
0.6
0.4
0.2
0.0
21,600
22,000
22,400
22,800
23,200
23,600
24,000
24,400
24,800
25,200
S
ORF1a
ORF1b
S
M
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000

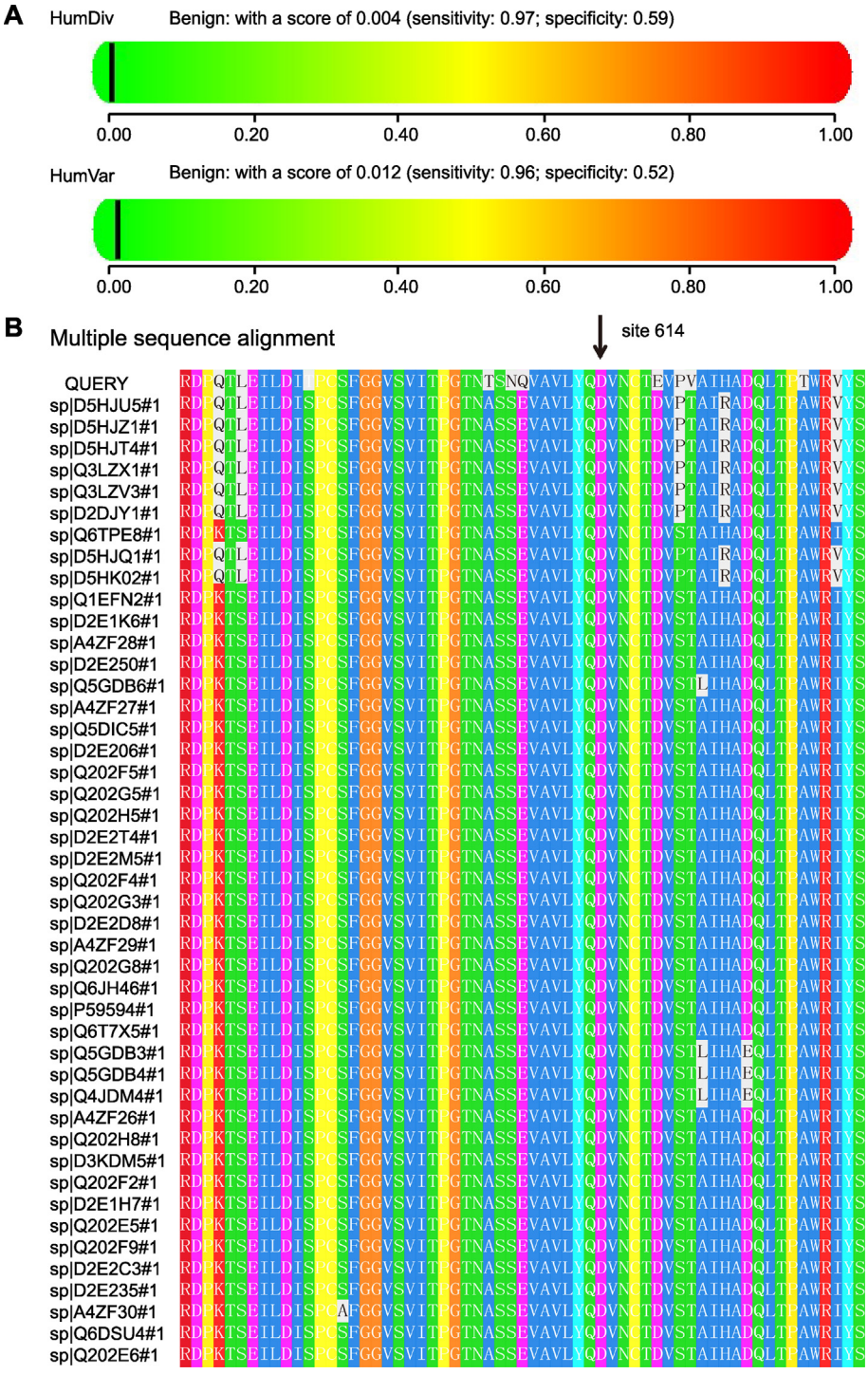


Figure 3 Variation in S protein at site 614 does not affect protein function. (A) Prediction results of the variation of S protein at 614. **(B)** Part multiple sequence alignment results, amino acids surrounding the variation position (614) were shown.

Figure 2 Amino acid variation of S protein at site 614 in Clade A2 and A2a SARS-Cov-2 strains. (A) Radial phylogenetic tree of 489 SARS-CoV-2 from Nextstrain, the cases were colored by the amino acid of S protein at site 614. Green: glutamic acid (D), yellow: glycine (G). Branch labels were clades. **(B)** Rectangular phylogenetic tree of 489 SARS-CoV-2 from Nextstrain, the cases were colored by the amino acid of S protein at site 614. Green: glutamic acid (D), yellow: glycine (G). Branch labels were the amino acid variation of SARS-Cov-2 proteins. **(C)** Diversity of S protein of SARS-Cov-2.

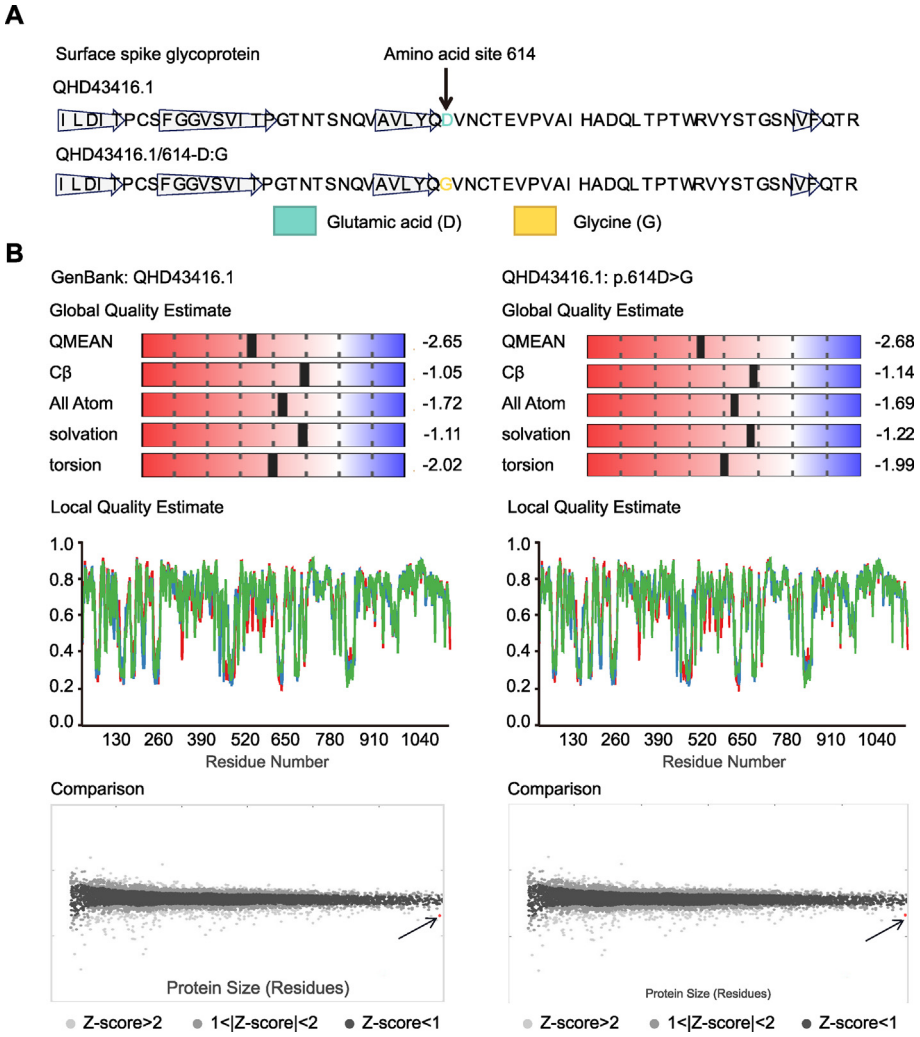


Figure 4 Protein modeling estimate of QHD43416.1 and QHD43416.1: p.614D > G. (A) Amino acids of QHD43416.1 and QHD43416.1: p.614D > G surrounding the site 614. **(B)** Protein modeling estimate results of QHD43416.1 and QHD43416.1: p.614D > G from SWISS-MODEL Server.

For example, studies on human immunodeficiency virus (HIV) evolution and replication in human cells identified a simian immunodeficiency virus (SIV) strain from chimpanzees (*Pan t. troglodytes*) as the progenitor of HIV-1.[33] Recently, via phylogenetic network analysis of SARS-Cov-2 and bat coronavirus genomes, Forster et al[34] identified three different types of virus (A, B, and C). Consistent with our findings, they stated that types A and C were only found outside East Asia, i.e., Europe and America, whereas type B was the most common type in East Asia. In addition, we found that 90% of patients with COVID-19 in China are aged over 20, consistent with that reported by the Chinese Center for Disease Control and Prevention.[2] As various studies have found an association between age and disease severity, reflecting the physiological and social changes of aging patients,[13,34–36] greater attention should be paid to aging patients infected with different types of SARS-CoV-2 variants.

It has been confirmed that SARS-CoV-2 uses the same cell entry receptor as SARS-CoV, i.e., via ACE2.[1,9] Specifically, the S protein of SARS-CoV-2 mediates receptor recognition of ACE2 and membrane fusion conformation in viral infection.[9] Thus, the S protein of SARS-CoV is a target for the development of vaccines.[27] We found that amino acid variation of the S protein at site 614 in SARS-Cov-2 (QHD43416.1: p.614D>G) contributed to Clades A2 and A2a. However, the virus strains in mainland China were not associated with this variation. In addition, using computational methods based on evolutionary principles to predict the effects of such variation on S protein function, we only identified natural or benign effects on function, with "does not affect," and "does not reduce" classified as neutral effects.[20] However, protein modeling[23] of the two proteins showed different scores in Cβ interaction, all-atom interaction, solvation, and torsion, thus highlighting different characteristics of the two proteins.[37–39] It has been reported that increased synonymous substitution rates of S proteins are caused by higher mutation rates.[40] In addition, the 3D structure of the proteins indicates that the S protein of SARS-CoV-2 shows higher binding affinity than SARS-CoV to ACE2.[12] Therefore, we believe that the mutant S protein with amino acid variation at site 614 more likely

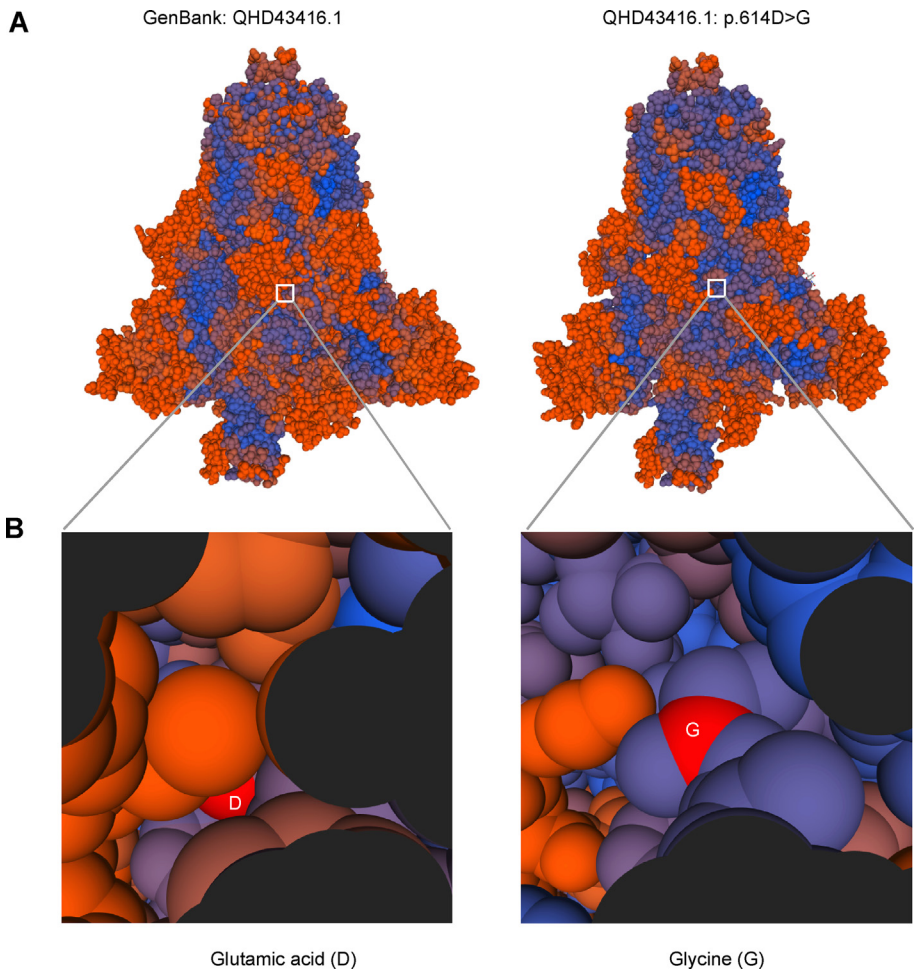


Figure 5 Three-dimensional (3D) protein structure models of QHD43416.1 and QHD43416.1: p.614D > G. (A) 3D protein structure of QHD43416.1 and QHD43416.1: p.614D > G performed by SWISS-MODEL Server. **(B)** The zooming-in region of site 614 of QHD43416.1 and QHD43416.1: p.614D > G performed by SWISS-MODEL Server.

contributes to viral infection; however, this requires further examination.

It should be noted, however, that we only explored differences in the phylogenetic tree and some epidemiological characteristics between mainland China and Clades A2 and A2a, and we do not know whether the variant was due to virus evolution or virus mutation. Indeed, although SARS-CoV-2, SARS-CoV, and MERS-CoV are naturally hosted in bats,[40] the specific transmission pathway from bats to humans remains to be determined. In addition, although the mutant S protein was predicted to have a neutral effect on protein function, different quality scores and 3D structures from the S protein were found, and thus different specificity of virus infection should be further analyzed. Furthermore, the analyzed SARS-CoV-2 strains from European regions were not ideally distributed, as a large cohort of cases were from the Netherlands, with two global hotspots, i.e., Italy and Spain, only accounting for a total of six cases. One reason may be that there were fewer infections in Italy and Spain in early March. As such, we re-explored the phylogenetic tree results in Nextstrain, and found that of the 37 and 149 cases in Italy and Spain, 35 (94.60%) and 72 (48.32%), respectively, were found in Clades A2 and A2a (Supplementary Figure 1), thus supporting our previous findings.

In summary, we found different genomic epidemiology among SARS-CoV-2 strains from mainland China in Clade B and Clade undefined and European strains in Clades A2 and A2a, specifically amino acid variation in the S protein at site 614 (QHD43416.1: p.614D>G). These results hint at potential viral genome divergence in SARS-CoV-2 strains.

Author contributions

C. C. performed the bioinformatics analyses and wrote the manuscript. P. W. and G. L. designed the study. K. L., K. M., L. H., Y. T., Y. Q., H. S., and W. D. provided critical feedback on the experiments and worked for discussions and coordination of the project.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the research-oriented clinician funding program of Tongji Medical College, Huazhong University of Science and Technology. The mentioned funding institutions played no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript. We are grateful for the computational analysis tools of Nextstrain, SWISS-MODEL, PROVEAN (v1.1) and PolyPhen-2. We sincerely thank the work done by all anti-epidemic personnel.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2020.05.006.

References

1. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565—574.
2. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese center for disease Control and prevention. *JAMA*. 2020;323(13):1239—1242. https://doi.org/10.1001/jama.2020.2648.
3. Del Rio C, Malani PN. COVID-19-New insights on a rapidly changing epidemic. *JAMA*. 2020;323(14):1339—1340. https://doi.org/10.1001/jama.2020.3072.
4. World Health Organization Coronavirus disease. 2019 (COVID-19) situation report — 101. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200430-sitrep-101-covid-19.pdf?sfvrsn=2ba4e093_2.
5. Li W, Shi Z, Yu M, et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science*. 2005;310(5748):676—679.
6. Ge XY, Li JL, Yang XL, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. 2013;503(7477):535—538.
7. Yang L, Wu Z, Ren X, et al. Novel SARS-like betacoronaviruses in bats, China. *Emerg Infect Dis*. 2011;19(6):989—991, 2013.
8. Hu B, Zeng LP, Yang XL, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog*. 2017;13(11), e1006698.
9. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270—273.
10. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science*. 2020;367(6485):1444—1448. https://doi.org/10.1126/science.abb2762.
11. Simmons G, Reeves JD, Rennekamp AJ, Amberg SM, Piefer AJ, Bates P. Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry. *Proc Natl Acad Sci USA*. 2004;101(12):4240—4245.
12. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog*. 2018;14(8), e1007236.
13. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497—506.
14. Yu W, Tang G, Zhang L, Corlett R. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool Res*. 2020;41(3):247—257. https://doi.org/10.24272/j.issn.2095-8137.2020.022.
15. Calisher C, Carroll D, Colwell R, et al. Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. *Lancet*. 2020;395(10226):e42—e43.
16. Kupferschmidt K. Genome analyses help track coronavirus' moves. *Science*. 2020;367(6483):1176—1177.
17. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121—4123.
18. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018;(1):4. vex042.
19. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265—269.
20. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PloS One*. 2012;7(10), e46688.
21. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745—2747.
22. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248—249.
23. Bienert S, Waterhouse A, de Beer TA, et al. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res*. 2017;45(D1):D313—D319.
24. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296—W303.
25. Eickmann M, Becker S, Klenk HD, et al. Phylogeny of the SARS coronavirus. *Science*. 2003;302(5650):1504—1505.
26. Luk HKH, Li X, Fung J, Lau SKP, Woo PCY. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect Genet Evol*. 2019;71:21—30.
27. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV—a target for vaccine and therapeutic development. *Nat Rev Microbiol*. 2009;7(3):226—236.
28. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. *Genome Biol*. 2003;4(11):R72.
29. Conrad B, Antonarakis SE. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genom Hum Genet*. 2007;8:17—35.
30. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 2011;27(3):343—350.
31. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223—230.
32. Cohen J. New coronavirus threat galvanizes scientists. *Science*. 2020;367(6477):492—493.
33. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med*. 2011;(1):1. a006841.
34. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA*. 2020;117(17):9241—9243. https://doi.org/10.1073/pnas.2004999117.
35. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507—513.
36. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in wuhan, China. *JAMA*. 2020;323(11):1061—1069. https://doi.org/10.1001/jama.2020.1585.

37. Fitzgerald JE, Jha AK, Colubri A, Sosnick TR, Freed KF. Reduced C(beta) statistical potentials can outperform all-atom potentials in decoy identification. *Protein Sci*. 2007;16(10):2123—2139.
38. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theor Comput*. 2017;13(6):3031—3048.
39. Zgarbova M, Luque FJ, Sponer J, Otyepka M, Jurecka P. A novel approach for deriving force field torsion angle parameters accounting for conformation-dependent solvation effects. *J Chem Theor Comput*. 2012;8(9):3232—3242.
40. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*. 2019;17(3):181—192.